Local Fitting of Regression Models by Likelihood: What's Important?

by

N.L. Hjort and M.C. Jones

Also included here are two bonuses for our readers:

READER COMMENT by M.C. JONES and N.L. HJORT on the paper BROWN, L.D. and HWANG, J.T.G. (1993), "How to Approximate a Histogram by a Normal Density", *THE AMERICAN STATISTICIAN* 47, 251–255

as well as Nils Lid Hjort's paper

MINIMUM L2 AND KULLBACK–LEIBLER ESTIMATION

which also appears in *Transactions of the 12th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes,* held from 29 August to 3 September, 1994

Local Fitting of Regression Models by Likelihood: What's Important?

M.C. JONES and N.L. HJORT

In this short essay, we look at an attractive way of performing semiparametric regression. A 'vehicle' parametric model is fit locally using kernel weights, in an extension of earlier local likelihood and local least squares polynomial fitting methods. We argue that performance is primarily affected by the number of parameters in the vehicle model for the regression mean, which determines the order of the bias. Secondly, there is also an effect of the exact form of the vehicle model: if this model is 'right' the method has the potential to behave in a fully parametric way, with its efficiency advantages, and otherwise the method should behave like a nonparametric estimator. Specifications like the right parametric form for the likelihood or perhaps just for the variance are relatively unimportant. These conclusions are based on asymptotic approximations.

KEY WORDS: Bias reduction; Kernel smoothing; Local likelihood; Loess; Nonparametric regression; Semiparametric regression.

M.C. Jones is Reader in Statistical Science, Department of Statistics, The Open University, Milton Keynes, MK7 6AA, U.K. N.L. Hjort is Professor of Statistics, Department of Mathematics and Statistics, University of Oslo, N–0316 Oslo, Norway. This paper was initiated when the first author visited Oslo in March 1994.

1. INTRODUCTION

"Local likelihood estimation" is a term coined by R.J. Tibshirani (see Tibshirani and Hastie, 1987) for a method to allow smoothing in quite general regression models. The methodology requires the specification of a conditional density function, that of a response Y conditioned on covariates X, which yields the likelihood function, and the *local* fitting of its mean to data. Localisation was carried out by Tibshirani and Hastie by a nearest neighbour weighting scheme which includes in the likelihood fitting only those (X, Y) pairs for which X is closest to estimation point x, and by approximating the regression mean function by a straight line in that neighbourhood. In this paper, we will discuss an extended version of these ideas with particular regard to replacing the line that is fitted locally with a more general model. This raises the following question which we attempt to answer in this paper. Which of the elements of this extended smoothing method, the likelihood function itself, the model chosen for local fitting, or some other detailed aspects of the method, are most important in the sense of having most impact on the quality of regression estimation?

First, though, why consider extending Tibshirani and Hastie's (1987) methodology at all? Tibshirani and Hastie note that as the degree of localisation is reduced, their method tends to a full likelihood fit of the linear model. Similarly, if a different model is employed, the limiting case as the amount of smoothing gets large will be a full likelihood fit of the chosen model. This limiting case is, of course, a standard parametric approach to regression. And as it occurs as a limit of what might have been thought of as nonparametric regression, the local likelihood method neatly becomes a semiparametric method. If one's 'vehicle' model (for the mean) happens to be a good global model for (the mean of) the data, then one can hope that little localisation will take place, and the efficiency advantages of parametric fitting will be available. But if the vehicle model is inappropriate, much localisation should happen, and the method thus will retain the flexibility advantages of the nonparametric approach. The idea can be found in special case form — e.g. fit a logistic function to binary data locally — in several places in the literature. See, for example, Kozek (1992) and Gozalo and Linton (1994).

Tibshirani and Hastie's (1987) local likelihood linear fitting is itself a generalisation of local least squares linear fitting, a method that was al-

ready popular with many through Cleveland's (1979) use of it in his "loess" smoother, and which has become even more popular because of the theoretical work of Fan (1992). See, for example, Hastie and Loader (1993). Least squares corresponds, of course, to likelihood fitting using constant variance normal conditional densities. We will follow Fan in replacing nearest neighbour localisation by kernel localisation in which the local model is fit using weights which attach greatest influence to points with X values at and very close to x, less weight to points a little way away from x, and no weight (or virtually none) to points far from x according to a kernel function K. Write $K_h(u) = h^{-1}K(h^{-1}u)$ and take K to be a symmetric nonnegative function with finite integral. The parameter h, the *bandwidth*, controls the degree of smoothing or localisation applied to the data, and clearly has a great effect on the resulting estimate. The resulting extended local likelihood method chooses the parameters θ of the vehicle model g as

$$\hat{\theta} = \hat{\theta}(x) = \arg \max_{\theta} \sum_{i=1}^{n} K_h(x - X_i)\ell\{Y_i, g(X_i; \theta)\} \tag{1}$$

and uses as smooth $g(x; \hat{\theta}(x))$. Here ℓ is the assumed log likelihood and $\{(X_i, Y_i), i = 1, \ldots, n\}$ is the data. For example, one might take $Y_i \sim N(\exp(-\theta X_i), \sigma^2)\}$ locally; this leads to minimisation of $\sum_{i=1}^{n} K_h(x-X_i)\{Y_i - \exp(-\theta X_i)\}^2$ for each x. This special case is inspired by Ansley, Kohn and Wong (1993) who took a spline approach to essentially the same problem. Another example might consider count variables Y_i to be Poisson with parameter $\exp(\theta_1 + \theta_2 X_i + \theta_3 X_i^2)$. (Notice that in cases like the normal example one could reasonably have at least one extra parameter related to the (local) variance, but we will not consider this in this paper.)

Let m be the true regression mean function and f an appropriately defined "design" or covariate density, the marginal density of X. Then Fan (1992) showed that the mean squared error of kernel weighted local least squares *linear* fitting, conditional on $X_1, \ldots, X_n$, asymptotically takes the attractive form

$$\frac{1}{4}h^4\mu_2^2(K)m''^2(x) + \frac{1}{nh}\frac{\sigma^2(x)}{f(x)}R(K) \tag{2}$$

as $n \to \infty$, $h = h(n) \to 0$ and $nh \to \infty$. Here $\mu_2(K) = \int u^2 K(u)\,du$, $R(K) = \int K^2(u)\,du$ and $\sigma^2(x) = \mathrm{var}(Y \mid X = x)$. The first term in (2) is the (asymptotic) squared bias and the second term is the (asymptotic) variance.

The bias term is appropriate provided m is sufficiently differentiable (it is not our intention to be specific about regularity conditions). For non-uniform f it is not all that trivial a task to obtain a bias proportional simply to $m''(x)$, a measure of the "roughness" of m, at least not whilst retaining the variance above (e.g. Jones, Davies and Park, 1994). Fan and Gijbels (1992) went on to show that local least squares linear fitting also achieves $O(h^2)$ bias at and near boundaries of the support of f (expression (2) assumes x is in the "interior" of the design space), and these facts combined with local least squares linear fitting's use of just a single bandwidth make it, or a closely related method to follow, the method of choice for many.

Now, is it the normality or the linearity or something else that is most crucial to achieving (2)? Fitting lines locally is clearly just a special case of fitting polynomials locally, and properties of the latter are well-understood (Fan and Gijbels, 1995, Ruppert and Wand, 1995); these have great relevance to what follows. Fitting degree zero polynomials (precisely the well-known Nadaraya–Watson estimator) results in asymptotic bias proportional to $h^2\{m''(x) + m'(x)f'(x)/f(x)\}$; degree two and three polynomials afford bias of order h^4, an improvement, degree three allowing that bias to depend only on $m''''(x)$, degree two involving $m'''(x)$ and the design density as well as $m''''(x)$ in its bias; order h^6 bias is achieved by fitting polynomials of degrees four and five with quintic polynomials having the simpler bias; and so on. (Local least squares cubic fitting is thus the popular alternative to local linear fitting alluded to above.) Boundary bias remains of order h^{2k}, say, for fitted polynomials of degree $2k-1$, but is $O(h^{2k-2})$ for degree $2k-2$. Non-normal likelihoods (together with local polynomial models) have been considered by Fan, Heckman and Wand (1995), and comparatively little dependence on the likelihood has been observed. So does the local model hold the key?

In Section 2, we argue that the vehicle model is indeed of the greatest importance, and in a rather interesting way. Our observations parallel some we made elsewhere (Hjort and Jones, 1995) in relation to local likelihood *density* estimation. Indeed, our purpose here is not to present further detailed theoretical results (nothing will be proved although much in this paper is novel) but rather to provide an informative essay on the local fitting of regression models by likelihood which elucidates the main structure. In a sense, our definition of the "most important aspects" of general local fitting is those that most affect *asymptotic* properties of the method. That is not to say, of course, that in any given small sample situation, other aspects of

the methodology might not also play a non-negligible role.

Our concern is with the case of a single global bandwidth h; we briefly note how further elements become important when using local bandwidths $h(x)$ in Section 3, as well as briefly discussing automatic global bandwidth choice. We deal with a univariate covariate throughout for clarity and simplicity. However, multiple covariates are important in practice and one can hope that these new methods will become particularly fruitful in higher dimensions.

2. IMPORTANCE OF THE VEHICLE MODEL

A quite general vehicle model for the mean affects the asymptotic bias in two ways. The more important of the two ways is the one capable of changing the order of the bias. And the order of the bias is determined solely by the *number of parameters* fitted in the vehicle model for the mean and not by more specific aspects of the parametrisation. One and two parameters yield bias of $O(h^2)$; two parameters afford the simple bias reliant only on second derivatives, one parameter results in reliance on first and second derivatives and the design density. Three and four parameters give bias of order h^4, four parameters yielding the simpler dependence on fourth derivatives only. Five and six parameters yield $O(h^6)$ bias, six parameters the more attractively of the two. And so on. Boundary bias is of order h^p where p is the number of parameters. (In some cases, one might have to take care about the active number of parameters in a naively parametrised vehicle mean model.) The parallel with the properties of local polynomial fitting spelt out in Section 1 is not, of course, accidental. This behaviour is a consequence of the ability to approximate smooth vehicle models locally by polynomials, thanks to Taylor's theorem.

But there are noteworthy differences even in asymptotic bias between vehicle models with the same number of parameters. This shows up in the derivative terms which were not spelled out in the previous paragraph. When we talked of bias depending on the k'th derivative, say, we actually meant on

$$\{m(x) - g_0(x)\}^{(k)} \tag{3}$$

where g_0 denotes the (locally) "best fitting" vehicle model $g(x, \theta_0(x))$ *to the true mean* m (where $\theta_0(x)$ represents the "best fitting" local parameters), and $g_0^{(k)}(x) \equiv \frac{\partial}{\partial z} g(x,z)|_{z=\theta_0(x)}$. The meaning of best fitting is defined by the

local kernel smoothed likelihood; formally, $\theta_0(x)$ is the parameter maximising the expected value of $K_h(x - X)\ell(Y, g(X;\theta))$ under the true distribution of (X, Y). The semiparametric nature of general local likelihood fitting is thus reflected: if g is the "correct" form of model, the discrepancy between truth and best parametric approximation will be zero and the leading terms in the bias will be zero. Indeed the theory then tells us that it will be optimal to increase the bandwidth without limit and hence to prefer a single global parametric fit with its efficiency advantage. (All that will remain is any bias in the (local) parametric estimation step, and, using a large bandwidth this should be of order n^{-1}.) If the vehicle model is not such a fortuitous choice, the bias should nonetheless be reduced when our choice is "close" to being an appropriate one, and may well not be inferior to the "nonparametric bias" comprising just $m^{(k)}(x)$ — really the bias of appropriate local polynomial fitting — even when our vehicle model has little in common with the true regression mean. Parametric behaviour when parametric is appropriate and nonparametric otherwise, the hallmark of good semiparametric estimation, is thus observed.

Aside. Hjort and Jones (1995) make the same points in local likelihood density estimation, but the interpretation for the regression case is new. Loader (1994) considers the attractive special case where the log density is modelled (locally) by polynomials, perhaps the most natural analogue of local polynomial fitting for density estimation. Copas (1994) has a slightly different definition of local likelihood for density estimation but has very much the same semiparametric outlook; it seems that the Copas and Hjort and Jones approaches are very similar asymptotically (see also Copas, 1995). For local *Bayesian* versions of the methodology discussed here, see Hjort (1995a) for density estimation and Hjort (1995b) for semiparametric regression.

An interesting class of examples that can be thought of as local likelihood fitting of certain vehicle models to the mean has been investigated by Fan, Heckman and Wand (1995). Fan et al. work with exponential family likelihoods where the mean parameter μ is related to a transformed parameter η — often η would be the canonical parameter — by a known link function, $\eta = g(\mu)$. Fan et al. extend generalised linear model ideas to the smoothing context by fitting polynomials $p(x)$ to η by kernel weighted exponential family likelihood. To tie in this work with the current, we can interpret this approach as the local likelihood fitting of vehicle models of the form $g^{-1}(p(x))$

to the mean. And sure enough in, for example, the case of $p(x)$ linear, Fan, Heckman and Wand's (1995) leading bias dependence (for mean estimation; their Theorem 2) on $\eta''(x)/g'(m(x)) = m''(x) + \{m'^2(x)g''(m(x))\}/g'(m(x))$ also arises from (3) by manipulating $m''(x) - \{\text{best fitting } g^{-1}(p(x))\}''$.

The work of Fan, Heckman and Wand (1995) also suggests that just which likelihood function is used is not so important, certainly relative to the vehicle model specification. In their wide class of models, the asymptotic variance of a local likelihood kernel estimator depends only on the (true, not assumed) $var(Y|X = x)$. In the case of even degree polynomials, the 'extra' bias term has some dependence on the *assumed* likelihood (see Fan et al.'s function ρ). Since in these cases, it is only the first two moments of the conditional density of Y given X that matter, Fan et al. further concentrate on the case of local *quasi-likelihood.*

But even more general local likelihoods can be considered, and the same type of behaviour observed. For instance, existing robust kernel M–estimation (e.g. Härdle and Gasser, 1984) can be thought of as assuming a convenient (local) likelihood based on a conditional density different from the one modelling most of the data. A general two parameter version of this would also have bias dependent only on $m''(x) - g_0''(x)$ and an appropriate term replacing the $\sigma^2(x)$ in the variance. (Härdle and Gasser actually worked with local constants, hence their $g_0''(x) = 0$, and with Gasser–Müller weights, which behave in bias terms like fitting local lines as explained by Jones, Davies and Park, 1994.) Fan, Hu and Truong (1992) explicitly deal with locally fitting lines using a general likelihood.

The message is that it is not very important which likelihood or quasi-likelihood one employs when using the attractive vehicle models with even numbers of parameters, and a choice might be made on simplicity or tractability grounds. (See also Jones's (1993a,b) discussion of the minor role of weighting for heteroscedasticity.)

Again, we must stress that 'importance' is being measured asymptotically, and one might well wish to be more circumspect for small samples. For example, the discussion of kernel M-estimation above concentrates on asymptotic efficiency and says nothing about small sample robustness properties!

3. ON BANDWIDTHS

Our discussion has throughout concentrated on a single global bandwidth choice for our otherwise local modelling. Authors often promote the use of a local bandwidth also, h replaced by $h(x)$, as done for local polynomial fitting by such as Cleveland (1979), Tibshirani and Hastie (1987) and Fan and Gijbels (1995). The final aspect of the previous section can then be added to the (asymptotically) important aspects of the methodology: one needs to employ a good estimate of the (local) variance function, and in this sense of the local likelihood, since the optimal local bandwidth depends on it. In fact, a handle on the design density is then needed for the same reason. Indeed, possible sparsity in the design is a prime reason for considering local bandwidth choice (e.g. Seifert and Gasser, 1994).

Reverting to consideration of a single global bandwidth h, good semiparametric performance depends on a good choice of h: to obtain the advantages of parametric fitting when the parametric model is appropriate requires the use of a large bandwidth, to ensure nonparametric behaviour when the vehicle model is inappropriate globally requires a smaller value of h. The asymptotics show this up, and also tell us that in the nonparametric case, the order of h should increase as the number of parameters does, e.g. $h \sim n^{-1/5}$ for one or two parameters, $h \sim n^{-1/9}$ for three or four parameters, and so on. Of course, this behaviour matches that of local polynomial fitting and, in turn, of the use of higher order kernels; moreover, increasing h with increasing p (to afford the extra information necessary for fitting more parameters locally) is natural intuitively. Data-based choice of optimal h is by no means straightforward, however. While plug-in methods, which directly estimate quantities in the asymptotic mean squared error, are effective, and have been developed for e.g. local polynomial fits by Ruppert, Sheather and Wand (1995), there is an extra complication in the semiparametric case: the estimation of terms of the form (3) rather than just $m^{(k)}$. An approach involving both the semiparametric estimator (to estimate $g_0^{(k)}$) and, say, a local polynomial special case (to estimate $m^{(k)}$) can be envisaged, but this remains to be worked out.

REFERENCES

Ansley, C.F., Kohn, R., and Wong, C.-M. (1993), "Nonparametric Spline Regression With Prior Information," *Biometrika*, 80, 75–88.

Cleveland, W.S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association,* 74, 829–836.

Copas, J.B. (1994), "Local Likelihood Based on Kernel Censoring," *Journal of the Royal Statistical Society,* Series B, to appear.

Copas, J.B. (1995), "Semi-Parametric Density Estimation by Likelihood," *Proceedings of the Kolmogorov Semester on Non-parametric and Semi-parametric Inference*, to appear.

Fan, J. (1992), "Design-Adaptive Nonparametric Regression," *Journal of the American Statistical Association,* 87, 998–1004.

Fan, J., and Gijbels, I. (1992), "Variable Bandwidth and Local Linear Regression Smoothers," *Annals of Statistics,* 20, 2008–2036.

Fan, J., and Gijbels, I. (1995), "Data-Driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation," *Journal of the Royal Statistical Society,* Series B, to appear.

Fan, J., Heckman, N.E., and Wand, M.P. (1995), "Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions," *Journal of the American Statistical Association*, to appear.

Fan, J., Hu, T.C., and Truong, Y.K. (1992), "Robust Nonparametric Function Estimation," Technical Report 035–92, Mathematical Sciences Research Institute, Berkeley.

Gozalo, P., and Linton, O. (1994), "Local Nonlinear Least Squares Estimation: Using Parametric Information Nonparametrically," Discussion Paper No. 1075, Cowles Foundation, Yale University.

Härdle, W., and Gasser, T. (1984), "Robust Non-Parametric Function Fitting," *Journal of the Royal Statistical Society,* Series B, 46, 42–51.

Hastie, T.J., and Loader, C. (1993), "Local Regression: Automatic Kernel Carpentry," (with comments) *Statistical Science*, 8, 120–143.

Hjort, N.L. (1995a), "Bayesian Approaches to Non- and Semiparametric Density Estimation," *Bayesian Statistics V*, to appear.

Hjort, N.L. (1995b), "Local Bayesian Regression," Statistical Research Report, University of Oslo.

Hjort, N.L., and Jones, M.C. (1995), "Locally Parametric Nonparametric

Density Estimation," Statistical Research Report, University of Oslo.

Jones, M.C. (1993a), "Do Not Weight for Heteroscedasticity in Nonparametric Regression," *Australian Journal of Statistics,* 35, 89–92.

Jones, M.C. (1993b), Contribution to the discussion of "Varying-coefficient models" by T. Hastie and R. Tibshirani. *Journal of the Royal Statistical Society*, Series B, 55, 791.

Jones, M.C., Davies, S.J., and Park, B.U. (1994), "Versions of Kernel-Type Regression Estimators," *Journal of the American Statistical Association,* 89, 825–32.

Kozek, A.S. (1992), "A New Nonparametric Estimation Method: Local and Nonlinear," in *Computing Science and Statistics. Proceedings of the 24th Symposium on the Interface*, ed. H.J. Newton, Fairfax, VA: Interface Foundation of North America, pp. 388–393.

Loader, C.R. (1994), "Local Likelihood Density Estimation," to appear.

Ruppert, D., Sheather, S.J., and Wand, M.P. (1995), "An Effective Bandwidth Selector For Local Least Squares Regression," *Journal of the American Statistical Association,* to appear.

Ruppert, D., and Wand, M.P. (1995), "Multivariate Locally Weighted Least Squares Regression," *Annals of Statistics,* to appear.

Seifert, B., and Gasser, T. (1994), "Infinite Variance For Local Polynomials: Analysis and Solutions," to appear.

Tibshirani, R.J., and Hastie, T. (1987), "Local Likelihood Estimation," *Journal of the American Statistical Association*, 82, 559–567.

BROWN, L.D., AND HWANG, J.T.G. (1993), "HOW TO APPROXIMATE A HISTOGRAM BY A NORMAL DENSITY", *THE AMERICAN STATISTICIAN* 47, 251–255: COMMENT BY JONES AND HJORT

Brown and Hwang (1993) consider fitting a normal distribution to data by minimizing "the integrated squared deviation between the histogram and the normal curve". The resulting parameter estimates depend to some extent on the histogram's bar width and, we would add, on the histogram's bar 'anchor'. By letting the bar width go to zero in Section 4 of their paper, Brown and Hwang eradicate both unnecessary dependences. All versions of this procedure, including the limiting one, are observed to provide robust estimates of mean and standard deviation.

The purpose of this letter is to point out that Brown and Hwang's limiting method is a special case of a quite novel and promising approach to robust estimation. Let f_θ denote the density of the distribution to be fitted to the data with θ representing its parameter(s). A natural strategy is to minimise an estimate of the L_2 distance $\int (f_\theta(x) - f(x))^2 dx$. Omitting the term that does not depend on θ and estimating $\int f_\theta(x) f(x) dx$ by $n^{-1} \sum_{i=1}^n f_\theta(x_i)$, this leads to the idea of choosing $\hat\theta$ to minimise

$$\int f_\theta^2(x) dx - 2n^{-1} \sum_{i=1}^n f_\theta(x_i). \tag{1}$$

That Brown and Hwang's Theorem 4.1 is precisely the unknown mean, unknown variance normal special case of the general method given by (1) is easy to verify.

Remarkably, we know of no earlier appearance of this natural approach in the literature even though there is much work in minimum distance estimation. Certain empirical transform methods are closest to it: working in Fourier space, minimum L_2 methods have been developed, but always with, effectively, a nonzero smoothing parameter. Although we have much experience with smoothing techniques, we remain firm believers in "don't smooth if you don't have to" and this may, after all, be a situation where you don't have to.

The robustness of estimates obtained from (1) is expected by the following heuristic reasoning. Maximum likelihood, via Kullback-Liebler deviation, is related to *weighted* integrated squared deviation with weight $1/f$

(for instance, compare f_θ with a smoothed $\hat{f}$ and do a Taylor expansion). This weighting causes maximum likelihood to take much more notice of the tails than does estimation by the unweighted case (1). So the unweighting, clearly, should aid robustness. (The robustness of minimum Hellinger distance methods deserves comment in the light of this. By the same token as above, Hellinger distance relates to $1/f_\theta$ weighted integrated squared deviation (hence good efficiency). Robustness then seems to be a consequence of smoothing (undesirably requiring choice of smoothing parameter). How do efficiency/robustness trade-offs compare?).

As usual, we sacrifice full efficiency for robustness. We have done some preliminary calculations. The influence function for this estimation scheme is not difficult to derive; it is typically bounded and in many cases redescending, hence robustness. Efficiency calculations for the normal case considered by Brown and Hwang are, however, a little disappointing: the asymptotic variances of Brown and Hwang's μ^* and σ^* are about 1.54 and 1.85 times those of the usual $\bar{x}$ and s, respectively. Perhaps this indicates that the direct minimum L_2 method is one that is quite robust but perhaps too inefficient.

We have, however, a further related idea that promises good robustness and better efficiency. It is a localised form of Kullback-Liebler fitting. In general terms a local kernel smoothed likelihood function yields an attractive semiparametric density estimation scheme, investigated in Hjort & Jones (1994). Taking this estimate at some well-chosen location gives robust parameter estimates. The smoothing parameter — yes, we have resorted to introducing smoothing! — now controls the robustness/efficiency tradeoff.

We each, independently, had (many of) these thoughts some time ago, but failed to prioritise the work highly enough to do anything further on it. We hope this will soon be remedied in collaboration with colleagues at the University of Texas at Austin.

M.C. JONES
Department of Statistics
The Open University
Milton Keynes, MK7 6AA
United Kingdom

N.L. HJORT
Department of Mathematics and Statistics
University of Oslo
PB 1053 Blindern, N–0316 Oslo 3
Norway

Reference

Hjort, N.L., and Jones, M.C. (1994), "Locally Parametric Nonparametric Density Estimation," in preparation.

MINIMUM L2 AND ROBUST KULLBACK–LEIBLER ESTIMATION*

Nils Lid Hjort, University of Oslo
Department of Mathematics, N–0316 Oslo, Norway

Abstract. *This paper introduces two new robust methods for estimation of parameters in a given parametric family. The first method is that of 'minimum weighted L2', effectively minimising an estimate of the integrated (and possibly weighted) squared error. The second is 'robust Kullback–Leibler', consisting of minimising a robust version of the empirical Kullback–Leibler distance, and can be viewed as a general robust modification of the maximum likelihood procedure. This second method is also related to recent local likelihood ideas for semiparametric density estimation. The methods are described, influence functions are found, as are formulae for asymptotic variances. In particular large-sample efficiencies are computed under the home turf conditions of the underlying parametric model. The methods and formulae are illustrated for the normal model.*

1. Minimum weighted L2 estimation. Let $X_1, \ldots, X_n$ be independent data points from an unknown density f, and suppose that the data are to be fitted to some given regular parametric family of densities $f_\theta(x)$. A simple and natural estimation idea is to minimise an estimate of $\int w(f_\theta - f)^2 \, dx$, where $w(.)$ is a suitable weight function, perhaps the constant 1. Disregarding the one term that does not depend on the parameter, this leads to the following strategy: minimise

$$Q_n(\theta) = \int w f_\theta^2 \, dx - 2\frac{1}{n}\sum_{i=1}^n w(x_i) f_\theta(x_i). \quad (1.1)$$

Taking the derivative this is also the same as solving

$$V_n(\theta) = \int w f_\theta u_\theta (dF_n - f_\theta \, dx) = 0, \quad (1.2)$$

where $u_\theta(x) = \partial \log f_\theta(x)/\partial\theta$ is the score function of the model, and where F_n is the empirical distribution of the data.

We derived (1.2) as a consequence of the natural (1.1), but forming an estimator by solving this second equation can also be motivated separately. It forces a weighted integral of the nonparametric $dF_n(x)$ to be equal to the corresponding weighted integral of the parametric $f_\theta(x)\,dx$. In spite of much work in the literature on various minimum distance strategies, the particular estimator (1.1)–(1.2) does not seem to have been studied earlier. It has also been proposed independently by M.C. Jones (personal communication). A method recently considered in Brown and Hwang (1993) has intentions similar to that of (1.1), but is unnecessarily hampered with an intermediate histogram approximation. This is a case of 'don't smooth if you don't have to'.

2. Influence function. Let $\widehat\theta$ be the estimator and let θ_0 minimise $\int w(f_\theta - f)^2 \, dx$. There is typically a unique parameter achieving this, and we interpret this θ_0 as the 'least false' or 'most appropriate' parameter value. As n grows $\widehat\theta$ converges almost surely to θ_0. Standard Taylor arguments show that

$$\widehat\theta - \theta_0 \doteq -V_n^*(\theta_0)^{-1} V_n(\theta_0), \quad (2.1)$$

where $V_n^*(\theta_0)$ is the matrix of derivatives of $V_n(\theta)$. Letting u_θ^* be the matrix of second order derivatives of the log density we have

$$V_n^*(\theta) = \int w(f_\theta^2 u_\theta u_\theta' + f_\theta u_\theta^*)(dF_n - f_\theta \, dx) - \int w f_\theta^2 u_\theta u_\theta' \, dx,$$

so that $-V_n^*(\theta_0) \to_p J = J(\theta_0)$, where

$$J(\theta) = \int w f_\theta^2 u_\theta u_\theta' \, dx - \int w(f_\theta^2 u_\theta u_\theta' + f_\theta u_\theta^*)(f - f_\theta) \, dx.$$

The influence function of the estimator can now be established, via (2.1); see for example Huber

From Proceedings of the 12th Prague Conference, 1994

(1981) for definition of and important uses of influence functions. Here it becomes

$$I(f,x) = J^{-1}\{w(x)f(x,\theta_0)u(x,\theta_0) - \xi_0\}, \quad (2.2)$$

where

$$\begin{aligned}\xi_0 &= \mathrm{E}_f w(X)f(X,\theta_0)u(X,\theta_0)\\ &= \int w(x)f(x,\theta_0)f(x)u(x,\theta_0)\,\mathrm{d}x\\ &= \int w(x)f(x,\theta_0)^2u(x,\theta_0)\,\mathrm{d}x.\end{aligned}$$

Where notationally convenient we write $f(x,\theta)$ for $f_\theta(x)$, and so on. The (2.2) function is typically bounded, which means robustness. The influence function is in fact also redescending in most cases, going to zero for x-values outside mainstream. This is often considered an attractive robustness feature of an estimation method.

3. Limit distribution. By the central limit theorem and the definition of θ_0, $\sqrt{n}V_n(\theta_0)$ tends to $\mathcal{N}\{0,M\}$, where

$$\begin{aligned}M &= \mathrm{VAR}_f\{w(X)f(X,\theta_0)u(X,\theta_0)\}\\ &= \int w^2 f_\theta^2 f u_\theta u_\theta'\,\mathrm{d}x - \xi_0\xi_0'.\end{aligned}$$

From (2.1) follows

$$\sqrt{n}(\widehat{\theta}-\theta_0) \to_d \mathcal{N}\{0, J^{-1}MJ^{-1}\}, \quad (3.1)$$

with J as given above. Note that this result has been reached without having to assume that the true f belongs to the parametric model.

The expressions for J and M simplify under model conditions. Of course there is some loss of efficiency, that is, the limiting covariance matrix $J^{-1}MJ^{-1}$ is larger than the best possible one under the model, namely $(\int f_\theta u_\theta u_\theta'\,\mathrm{d}x)^{-1}$, achieved by the maximum likelihood method.

4. Local and weighted L2 fitting. The size of the limiting variances depend on the weight function $w(.)$. Choosing a local weight function can be contemplated, say of the kernel type $K_h(x_0 - t)$ around a given x_0. Here $K_h(u) = h^{-1}K(h^{-1}u)$ and K is a given kernel function. This gives a locally estimated normal, for example, in a spirit similar to local likelihood methods discussed in Hjort and Jones (1994). The apparatus above can be used to investigate influence functions and large-sample properties.

It is sometimes desirable to let the weight function be data driven too, perhaps to increase precision under close to the model circumstances. One example would be to use $w_n(x) = w_0((x-\widetilde{\mu})/\widetilde{\sigma})$, for a suitable $w_0(.)$ function, with preliminary robust estimates of location and scale. Result (3.1) is still true under appropriate conditions, with J and M being defined in terms of the limit function version of $w_n(.)$.

5. Local and robust Kullback–Leibler fitting. The local kernel smoothed likelihood function, around a given x_0, is

$$\begin{aligned}L_n(x_0,\theta) = \sum_{i=1}^n K_h(x_i - x_0)\log f(x_i,\theta)\\ - n\int K_h(t-x_0)f(t,\theta)\,\mathrm{d}t,\end{aligned} \quad (5.1)$$

see Hjort and Jones (1994). As shown in Hjort and Jones (1994), maximising (5.1) aims at minimising the localised Kullback–Leibler distance

$$\begin{aligned}d(f,f_\theta) = \int K_h(t-x_0)\big[f(t)\log\frac{f(t)}{f_\theta(t)}\\ - \{f(t) - f_\theta(t)\}\big]\,\mathrm{d}t\end{aligned}$$

from true density to parametric density. In other words, the maximiser of (5.1) aims at a 'least false' parameter value θ_0 that in general is different from the one associated with the minimum weighted L2 method. Note that a large h gives a flat $K_h(t-x_0)$ function, and brings back the ordinary Kullback–Leibler distance and the traditional full likelihood method.

The aim of Hjort and Jones (1994) is primarily the complete semiparametric estimation of the full density curve, as partly opposed to concentrating on the locally estimated parameters themselves. But this is also automatically one way of obtaining robust parameter estimates for a given parametric family: Apply the above

for a suitable centrally placed x_0, for a reasonably sized h. The resulting maximiser $\widehat{\theta}$ is a robust estimate of θ, and $f(x,\widehat{\theta})$ a robust estimate of the underlying density curve.

Hjort and Jones (1994) demonstrate that

$$\sqrt{n}(\widehat{\theta}-\theta_0) \to_d \mathcal{N}\{0, J_h^{-1}M_hJ_h^{-1}\},$$

with certain generally valid expressions available there for J_h and M_h. At the moment it will suffice to give these under model conditions:

$$\begin{aligned} J_h &= \int K_h(t-x_0)u_\theta u'_\theta f_\theta \,\mathrm{d}t, \\ M_h &= \int K_h(t-x_0)^2 u_\theta u'_\theta f_\theta \,\mathrm{d}t - \xi_0\xi_0', \end{aligned} \tag{5.2}$$

where $\xi_0 = \int K_h(t-x_0)u_\theta f_\theta \,\mathrm{d}t$. The influence function of this robustified maximum likelihood is also derived in Hjort and Jones (1994), and is of the form

$$I(f,x) = J_h^{-1}\{K_h(x-x_0)u(x,\theta) - \xi_0\}. \tag{5.3}$$

This is reasonably similar to the influence function (2.2) for the minimum L2 method. In many cases the present method, with a suitably chosen h, is more efficient at the model than at least the unweighted version of the minimum L2 method.

6. The normal model. The most important special case is that of fitting data to a normal (μ,σ^2).

6.1. THE MINIMUM L2 METHOD. From (1.2) two equations are easily put up to define minimum L2 estimators $\widehat{\mu}$ and $\widehat{\sigma}$. These are solved for example by the iterative Newton–Raphson technique. Regarding performance, under Gaußian circumstances, and using a constant weight function, we find

$$\begin{aligned} J &= (\sigma^3\sqrt{2\pi})^{-1}\mathrm{diag}(1/2^{3/2}, 3/2^{5/2}), \\ M &= (\sigma^4 2\pi)^{-1}\mathrm{diag}(1/3^{3/2}, 2/3^{3/2} - 1/8). \end{aligned}$$

This gives an asymptotic variance for $\widehat{\mu}$ of size $1.5396\,\sigma^2/n$ and an asymptotic variance for $\widehat{\sigma}$ of size $0.9241\,\sigma^2/n$. These should be compared to the minimum possible values, under model conditions. These optimal figures are achieved by the ML method, and are σ^2/n and $\frac{1}{2}\sigma^2/n$, respectively. This makes the direct minimum L2 method qualify as a 'quite robust but perhaps too inefficient method'. Increased efficiency at the model is achieved through appropriate choices of weight function $w(.)$, cf. comments at the end of Section 4. One possibility here is $w_n(x) = \exp\{\frac{1}{2}\delta(x-\widetilde{\mu})^2/\widetilde{\sigma}^2\}$, defined in terms of preliminary robust estimates of location and scale, and with an extra tuning parameter $\delta \in (0,1)$. Choosing e.g. $\delta = 0.8$ leads to quite good efficiency at the model, while still retaining a reasonable robustness.

6.2. THE ROBUSTIFIED ML METHOD. The robust Kullback–Leibler fitting method of Section 5 can easily be made better than the unweighted minimum L2 method. For the present normal model, let us use a normal kernel. The method is then to minimise, for given x_0, the function

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n \phi\Big(\frac{x_i-x_0}{h}\Big)\frac{1}{h}\{\log\sigma + \tfrac{1}{2}(x_i-\mu)^2/\sigma^2\} \\ &\quad + \phi\Big(\frac{x_0-\mu}{\sqrt{\sigma^2+h^2}}\Big)\frac{1}{\sqrt{\sigma^2+h^2}} \end{aligned} \tag{6.1}$$

over all (μ,σ). We may compute the J_h and M_h matrices of (5.2) without serious difficulties. But in the present context the interest lies more in getting hold of a single, robust (μ,σ)-estimate, than in obtaining a full function of local estimates. Therefore we suggest using $x_0 = \widetilde{\mu}$, a robust preliminary estimate of the mean, say the simple median. Minimising (6.1) with this x_0 defines the proposed $\widehat{\mu}$ and $\widehat{\sigma}$.

Again it is of interest to see how well the method fares under Gaußian home-turf conditions. Somewhat arduous calculations give two diagonal matrices (J_μ, J_σ) and (M_μ, M_σ) for J_h and M_h of (5.2). Here

$$J_\mu = \frac{1}{\sigma^2}\frac{1}{\sqrt{2\pi}}\frac{1}{h}\frac{1}{R^3} \quad\text{and}\quad M_\mu = \frac{1}{\sigma^2}\frac{1}{2\pi}\frac{1}{h^2}\frac{1}{S^3},$$

in which

$$R = (1+\sigma^2/h^2)^{1/2} \quad\text{and}\quad S = (1+2\sigma^2/h^2)^{1/2}.$$

Similarly,

$$J_\sigma = \frac{1}{\sigma^2}\frac{1}{\sqrt{2\pi}}\frac{1}{h}\Big(1-\frac{2}{R^2}+\frac{3}{R^4}\Big),$$

$$M_\sigma = \frac{1}{\sigma^2}\frac{1}{2\pi}\frac{1}{h^2}\Big\{\frac{1}{S}\Big(1-\frac{2}{S^2}+\frac{3}{S^4}\Big) - \frac{1}{R^2}\Big(1-\frac{1}{R^2}\Big)^2\Big\}.$$

Thus $\sqrt{n}(\widehat{\mu}-\mu) \to_d \mathcal{N}\{0,\kappa_\mu^2\}$ and $\sqrt{n}(\widehat{\sigma}-\sigma) \to_d \mathcal{N}\{0,\kappa_\sigma^2\}$, where the asymptotic variances are found as M_μ/J_μ^2 and M_σ/J_σ^2. Some calculations give

$$\kappa_\mu^2 = \sigma^2\frac{R^6}{S^3} = \sigma^2\frac{(1+1/k^2)^3}{(1+2/k^2)^{3/2}},$$

writing $h = k\sigma$, and similarly

$$\kappa_\sigma^2 = \sigma^2\frac{(1+1/k^2)^2}{(1+2/k^2)^{5/2}} \frac{(1+1/k^2)^3(2+4/k^4)-(1+2/k^2)^{5/2}1/k^4}{(2+1/k^4)^2}.$$

How large should h be chosen? We think of h as $k\widetilde{\sigma}$, where $\widetilde{\sigma}$ is a robust preliminary estimate of standard deviation, and need to choose the factor k. As a mild surprise the value $k = 1$ gives precisely the same large-sample variances under the model as the straightforward minimum L2 method of Section 2, respectively $1.5396\,\sigma^2$ and $0.9241\,\sigma^2$. A more efficient but still quite robust value would be $k = 2$, 'place a normal with two standard deviations around the median and maximise the local kernel smoothed likelihood'. Then the values are $1.063\,\sigma^2$ and $0.563\,\sigma^2$, only a few percent above the values that are optimal under the model, viz. σ^2 and $\sigma^2/2$. Increasing the value to three estimated standard deviations brings the large-sample variances further down to $1.015\,\sigma^2$ and $0.5152\,\sigma^2$. One should not go much further if robustness is aimed for, but of course a large h gives back these optimal values.

Comparing the performance of the weighted L2 method, say with the data driven weight function indicated above, with that of the robust Kullback–Leibler estimator, is an interesting problem for further research. One should also devise criteria for choosing the necessary fine-tuning parameters.

7. Robust estimation of location and covariance matrix. The ideas and results above generalise easily to the multi-dimensional case. In particular the localised Kullback–Leibler method seems to constitute a fruitful way of obtaining robust estimates of μ and Σ, the mean vector and covariance matrix of the underlying distribution. The estimates can be viewed as robust estimates of these parameters under normality assumptions but also outside normality.

One concrete version of this scheme, in the p-dimensional case, is as follows: Start out with preliminary and robust estimates $\widetilde{\mu}$ and $\widetilde{\Sigma}$ for mean and covariance matrix. Then carry out local likelihood estimation with a Gaußian kernel function centred at $\widetilde{\mu}$ and with covariance matrix of size $h^2\widetilde{\Sigma}$. This is seen to be the same as minimising the criterion function

$$\Big[\frac{1}{n}\sum_{i=1}^n \frac{\exp\{-\frac{1}{2}(x_i-\widetilde{\mu})'\widetilde{\Sigma}^{-1}(x_i-\widetilde{\mu})/h^2\}}{h^p|\widetilde{\Sigma}|^{1/2}} \{\tfrac{1}{2}\log|\Sigma| + \tfrac{1}{2}(x_i-\mu)'\Sigma^{-1}(x_i-\mu)\}\Big] + \frac{\exp\{-\frac{1}{2}(\mu-\widetilde{\mu})'(h^2\widetilde{\Sigma}+\Sigma)^{-1}(\mu-\widetilde{\mu})\}}{|h^2\widetilde{\Sigma}+\Sigma|^{1/2}}$$

over all possible (μ,Σ). Note that this method properly generalises that of (6.1). For h larger than say 5 the procedure is practically the same as ordinary maximum likelihood estimation. A value of perhaps $h = 2$ constitutes a modified maximum likelihood procedure with quite good robustness qualities without sacrificing much in efficiency under multinormal conditions.

Acknowledgements. This paper has benefited from ongoing joint work on related matters with M.C. Jones, I.R. Harris and A. Basu.

References

Brown, L.D. and Hwang, J.T.G. (1993). How to approximate a histogram by a normal density. *American Statistician* **47**, 251–255.

Hjort, N.L. and Jones, M.C. (1994). Locally parametric nonparametric density estimation. Submitted for publication.

Huber, P.J. (1981). *Robust Statistics.* Wiley, New York.